Le dictionnaire des francophones, une plateforme lexicographique contributive et sémantique

F Limpens, Nicolas Delaforge, P-R Lhérisson, N Gasparini

Le dictionnaire des francophones, une plateforme lexicographique contributive et sémantique.

F. Limpens[1], Nicolas Delaforge[1], P-R. Lhérisson[1], N. Gasparini[2]

[1]SCIC Mnémotix, https://mnemotix.com
[2]Institut international pour la Francophonie, Univ. Jean Moulin Lyon 3 (2IF)

freddy.limpens@mnemotix.com

Résumé

Le Dictionnaire des francophones (DDF) est un projet inédit de plateforme articulant ressources lexicographiques savantes et issues de contributions de la communauté. Ces différentes ressources prennent la forme d'un graphe de connaissances rendu accessible en lecture et/ou écriture par une architecture innovante alliant l'état de l'art des technologies du Web sémantique et du big data. Cet article propose une visite guidée de cette plateforme aujourd'hui accessible en ligne et qui illustre certaines des problématiques typiques de l'Ingénierie des Connaissances: l'intégration et l'interopérabilité de sources de données hétérogènes, et la gestion des cycles de vie de ces mêmes ressources, incluant la gestion des contributions et des droits d'accès sur un graphe RDF. L'article se conclut sur les perspectives et les prochaines étapes de ce projet s'inscrivant dans l'écosystème de l'Open Linked Data et des communs logiciels.

Mots-clés

Lexicographie, Dictionnaires, Web Sémantique, Ontolex, Approche Contributive

Abstract

The Dictionnaire des francophones (DDF) is a novel project that articulates lexicographic resources coming from scholars and from community contributions. These different resources take the form of a knowledge graph that evolves thanks to an innovative platform architecture combining state of the arts semantic web and big data technologies. This article proposes a guided tour of this platform, available online, and which illustrates some of the typical problems of Knowledge Engineering: the integration and interoperability of heterogeneous data sources, and the management of the life cycles of these same resources, including the management of contributions and access rights on an RDF graph. The article concludes with the prospects and next steps for this project set in the Open Linked Data and Software Commons ecosystem.

Keywords

Lexicography, Dictionaries, Semantic Web, Ontolex, Contributive approaches

Introduction

Le Dictionnaire des francophones (DDF) est une base d'informations sur les mots du français dont l'interface principale est celle d'un dictionnaire de définitions avec leurs aires d'usage et bien d'autres informations. Mais le DDF se pose avant tout comme une nouvelle ressource lexicographique contributive visant à compléter et prolonger des approches comme le Wiktionnaire[1] en ouvrant davantage encore le champ des contributions possibles. Le DDF décrit la richesse et la diversité du français parlé au sein de l'ensemble de l'espace francophone. C'est un projet institutionnel et académique: impulsé par le gouvernement français en mars 2018, ce projet a été transformé en actes par différentes institutions[2] dont l'Institut international pour la Francophonie (2IF) qui en est l'opérateur. La société Mnémotix[3] s'en est vu quant à elle confier le développement. Le contenu initial du DDF est issu de travaux existants et il s'enrichira grâce à l'implication du lectorat dans la description des usages.

Dans cet article nous nous attachons à montrer en quoi l'approche de l'Ingénierie des Connaissances combinées aux standards du Web Sémantique et à une architecture réactive et taillée pour traiter de gros flux de données ont permis de répondre aux défis posés par le projet de la plateforme DDF. Cette plateforme sera lancée officiellement et accessible au public le 16 mars 2021[4].

Modèle de données et approche contributive

Une des contributions majeures du DDF est d'offrir un accès unique, sous formes de données liées ouvertes[5] à un ensemble

[1] https://fr.wiktionary.org/
[2] La Délégation générale à la langue française et aux langues de France (DGLFLF), l'Institut international pour la Francophonie composante de l'Université Jean Moulin Lyon 3 (2IF), l'Organisation internationale de la Francophonie (OIF) et l'Agence universitaire de la Francophonie (AUF)
[3] https://mnemotix.com
[4] https://www.dictionnairedesfrancophones.org/
[5] i.e. adoptant des préceptes du Linked Open Data, et offrant notamment un accès SPARQL aux données du DDF.

de ressources lexicographiques déjà existantes et en construction. Le Dictionnaire des francophones intègre ainsi: l'*Inventaire des particularités lexicales du français en Afrique noire (Inventaire),* le *Wiktionnaire* francophone, le *Dictionnaire des synonymes, des mots et expression du français parlé dans le monde (ASOM),* Le *Grand Dictionnaire terminologique (GDT)*, l'ouvrage *Belgicismes - Inventaire des particularités lexicales du français en Belgique,* le *Dictionnaire des régionalismes de France (DRF)* et la *Base de données lexicographiques panfrancophone (BDLP)*. *FranceTerme* est en cours d'intégration. Les trois premières ressources sont diffusées sous licence libre CC BY-SA 3.0 tandis que les quatre suivantes sont diffusées sous licence CC BY-SA-ND 4.0.

Toutes ces ressources ont pu être agrégées et liées entre elles en un graphe homogène de connaissances grâce à un modèle RDFS pivot, l'ontologie DDF[6] (Steffens et. al. 2020), basée sur l'ontologie Ontolex Lemon (McCrae et al. 2017). Dans ce modèle, les entrées d'un dictionnaire s'articulent autour d'unités lexicales (`ontolex:LexicalEntry`) pouvant avoir plusieurs formes et plusieurs définitions (sens). Quelques ajustements ont dû être apportés à ce modèle pour affiner la granularité de description des propriétés lexicales. Par exemple, la propriété `ddf:place` permet de lier une forme ou une définition à une localité représentée par une URI geonames[7], et un ensemble de propriétés spécifiques permettent d'ajouter de nombreux marqueurs sémantiques aux définitions comme la connotation, le domaine, etc. Un ensemble de vocabulaires contrôlés (au format SKOS) permettent par ailleurs de structurer tous ces descripteurs. Enfin les relations sémantiques entre formes et (ou entre) définitions sont également décrites de manière complète, via une réification RDF et un vocabulaire contrôlé dédié.

Aux dictionnaires importés s'ajoute une ressource lexicographique contributive qui suit le même modèle de données. Le DDF s'inscrit ainsi dans l'approche de la lexicographie contributive, riche en initiatives variées (Dolar, 2017). Le modèle contributif du DDF repose sur une approche additive et non agrégative, c'est-à-dire que chaque contributrice ou contributeur peut soit ajouter une nouvelle forme, ou une nouvelle définition à une forme existante, ou ajouter une nouvelle information (exemple, marqueur de définition, etc.) à une définition ou forme existante. Les contributions se complètent et complètent également les ressources importées, et n'ont pas le devoir de fusionner. Cette approche résolument ouverte pose certaines contraintes et a orienté profondément la conception de la plateforme, que nous décrivons dans ce qui suit.

Conception et implémentation de la plateforme

Intégration des dictionnaires existants

Les données du DDF proviennent de dictionnaires sérialisés dans des formats hétérogènes, et pour la plupart sans point d'accès (API). Pour chacune de ces différentes ressources nous avons établis une correspondance vers le modèle pivot (cf supra). En fonction de leur nature nous avons traité ces données en adoptant des stratégies différentes. Nous avons mis en place un traitement génériques pour les dictionnaires qui comptent moins de dix mille entrées et un traitement spécifique pour le wiktionnaire français qui compte plus de 1844360 entrées[8] (en incluant les flexions).

Pour les dictionnaires qui contiennent peu d'entrées, nous avons utilisé un modèle pivot que nous avons décliné en XML et en CSV. Le modèle XML a été utilisé principalement dans le cas de l'Inventaire. Nous avons extrait le contenu du dictionnaire d'un fichier PDF et nous l'avons converti en XML en suivant le modèle OWL. Nous avions choisi cette stratégie afin de permettre aux experts linguistes d'intervenir sur la source et avoir ainsi une conversion RDF optimale.

Pour l'insertion des autres dictionnaires (GDT, BDLP, ASOM, DRF) nous avons opté pour le format CSV beaucoup plus facilement lisible et modifiable à l'aide d'un tableur. Les dictionnaires à insérer sont convertis en un fichier CSV en suivant un modèle strict ou chaque colonne du fichier correspond à une classe ou une propriété OWL du modèle DDF. Un parseur lit ce fichier et produit le RDF. Ce modèle pivot permet de rendre générique la création des triplets et d'être agnostique aux formats sources des dictionnaires.

Pour convertir le Wiktionnaire en RDF nous avons adopté une stratégie différente. Des dumps du Wiktionnaire sont disponibles dans un format XML, ce qui facilite l'accès à la donnée. Le contenu de chaque page du Wiktionnaire est quant à lui encodé dans un format texte appelé wikicode dont la syntaxe est définie par une documentation consultable sur le site de mediawiki. De par la nature collaborative du Wiktionnaire, cette syntaxe n'est pas toujours suivie de manière rigoureuse, ce qui rend difficile l'écriture d'un parseur universel permettant de lire le wikicode (Navarro et al., 2009; Sajous et al., 2010, 2011). Nous avons écrit un parseur pour sélectionner les entrées françaises, découper le wikicode de ces entrées selon des sections, des définitions, des exemples et des citations, et pour parser les modèles du wikicode afin qu'ils aient un rendu textuel lisible. Nous avons pu par la suite transformer en RDF les données du Wiktionnaire, ce qui constitue une des contributions notables de ce projet, pour le contenu en français du Wiktionnaire, complétant ainsi d'autres rares approches similaires comme DBnary (Serasset, 2015) pour le contenu multilingue.

Chaque Dictionnaire transformé en RDF est intégré dans un graphe nommé. Cela a pour but de faciliter la gestion des licences des dictionnaires et de permettre des mises à jour séparées. Le schéma de nommage des URIs prend ici une importance capitale et doit être déterministe afin que les URIs des ressources importées ne changent pas entre les différents imports tant que leur contenu reste identique. Ceci permet en retour que les liens pointant vers ces URIs ne soient pas cassés à chaque nouvel import. Ainsi les URIs des `lexicog:Entry` ou des `ontolex:LexicalEntry` ne changent pas, en

[6] https://gitlab.com/mnemotix/ddf/ddf-models
[7] http://geonames.org/
[8] https://fr.wiktionary.org/wiki/Wiktionnaire:Statistiques

revanche les URIs des définitions changent si le contenu de la définition change ou si des exemples et des citations sont ajoutés à la définition. Des routines calculant les différences seront par la suite développées afin de permettre un réalignement entre contenus ayant peu évolué, et ainsi préserver les contributions liées à ces contenus.

Conception de la plateforme

Ces dictionnaires importés sont stockés dans un triple store (GraphDB[9]) et servis par une plateforme se présentant comme une interface de navigation et de contribution dans un corpus de définitions. Afin de permettre un accès rapide, ces données sont indexées dans un moteur de recherche (Elasticsearch). La page d'accueil de la plateforme permet de saisir une forme et de voir toutes les définitions associés à cette forme. Si une forme ne retourne aucun résultat, par exemple si elle a été mal orthographiée, le moteur de recherche suggère des formes similaires.

La contribution au DDF peut se faire pour l'instant par l'ajout de définitions avec ou sans nouvelle forme. Ainsi, une des premières difficultés qui se pose est de permettre un tri parmi l'ensemble des définitions proposées, contributives ou importées. Par exemple, le mot "faire" en compte plus de 83 sans compter les définitions contributives. L'approche ici adoptée est d'établir un tri dans la liste des résultats selon de multiples critères:

- la localisation choisie par l'utilisatrice depuis la page d'accueil: sont affichés en premières les définitions liées à une localité la plus proche.
- score issu de la validation ou du signalement par le lectorat: sur la page de chaque définition, les usagers peuvent voter ("valider") pour cette définition, ou bien également la signaler (en pouvant préciser le motif de signalement). Le comptage des validations (score positif) et des signalements (score négatif) permet d'établir un score global et les définitions ayant le score le plus haut remonteront dans la liste.
- score issu de la modération: la modération du contenu est assurée par une communauté de personnes membres du DDF et ayant le statut "opérateur"; celles-ci peuvent marquer certaines définitions comme "à supprimer". L'action de suppression est quant à elle effectuée par les administrateurs ou administratrices du DDF qui se voient suggérer automatiquement les définitions relevées par les membres "opérateurs" dans l'espace d'administration de la plateforme. Le statut "à supprimer" pour une définition contribue à faire baisser son score et donc sa place dans la liste des résultats.

La localisation des définitions est un critère de différenciation important dans le DDF, et celui-ci est mis en avant de manière graphique. Un code couleur correspondant aux continents d'origine des différentes localisations possibles permet de rapidement identifier la provenance d'une définition et oriente la lectrice ou le lecteur.

Le système de curation collective, via le vote ou le signalement de tous les contributeurs, et le système de modération *a posteriori,* via les actions des "opérateurs" et des administrateurs, sont les 2 aspects de la stratégie adoptée par les concepteurs du DDF pour permettre à la fois une grande liberté de contribution, et en même temps de mettre en avant les contenus de plus grande qualité et au contraire identifier et éliminer les contenus inappropriés.

D'un point de vue technique, la dimension contributive repose également sur une structuration des données basée sur les standards de Web Sémantique. L'ontologie pivot a donc été étendue[10] pour permettre d'outiller tous les aspects de la contribution (édition, droits d'accès), de la modération (profils d'usagers, signalements) et de la curation collaborative (vote). La modélisation des contributions du DDF est basée sur le modèle d'actions et de provenances PROV[11]. Ce modèle très simple mais très puissant repose sur une triade Entité-Action-Agent sur laquelle nous avons greffé le modèle de contribution du DDF. Les "Entités" sont ici toutes les contributions possibles du DDF (définition, forme, exemple, etc); les "Actions" réifiées sur ces entités (`:Creation`, `:Update`, `:Deletion`) permettent de versionner ces contributions (et par exemple d'annuler une suppression). Enfin les différents profils des contributeurs, et leurs droits d'accès correspondants, sont modélisés à partir d'une extension de SIOC[12] comme des groupes d'utilisateurs (`sioc:UserGroup`) auxquels sont rattachés les "Agents" (`sioc:UserAccount`).

Une architecture réactive et résiliente

Comme nous venons de le voir précédemment, l'ensemble des données importées et produites par les usagers est stocké au format RDF dans un triple store. Afin d'assurer le niveau de réactivité et de performance requis par une plateforme grand public telle que le DDF, nous avons mis au point une architecture reposant sur l'état de l'art des technologies du Web Sémantique et Big Data.

Le triple store choisi est GraphDB, proposé par Ontotext, et structuré ici en cluster afin d'assurer la résilience et une disponibilité optimale du système. De plus, l'accès en lecture des données est assuré par un moteur de recherche, en l'occurrence ElasticSearch, sur lequel les données sont indexées au chargement des données et de manière incrémentale; le graphe et les indexes sont ainsi synchronisés en permanence.

L'usage conjoint de ces 2 types de datastores (index et graphe) requiert cependant une architecture réactive que nous avons mise au point au sein de Mnémotix sous la forme d'un middleware générique, Synaptix[13]. Ainsi, chaque requête est traitée de manière asynchrone et non bloquante par un système de micro-services fédérés par un système de bus de messages implémentant le standard AMQP[14]. Chaque datastore est connecté au bus de messages par des "guichets" AMQP de type "Producer" ou "Consumer". Les "Consumers" ont des

[9] https://graphdb.ontotext.com/

[10] voir https://mnemotix.gitlab.io/ddf/ddf-models/
[11] https://www.w3.org/TR/prov-o/
[12] https://www.w3.org/Submission/sioc-spec/
[13] https://gitlab.com/mnemotix/synaptix
[14] https://www.amqp.org/

files où sont stockés les messages à traiter. Chaque message est aiguillé et traité individuellement jusqu'à ce que la file de messages soit vidée. Synaptix s'appuie sur ce système de messages pour synchroniser les datastores entre eux et pour ingérer des sources de données externes sans ralentir le fonctionnement global de la plate-forme.

Un autre aspect permettant des gains substantiels en termes de réactivité est que l'API publique n'est plus une API REST classique mais une API GraphQL qui permet de combiner plusieurs requêtes en un seul appel HTTP. Cette technologie développée par Facebook a été inventée pour répondre au problème du nombre important de requêtes qu'il fallait pour construire les murs de données des utilisateurs et qui saturait les serveurs Facebook. Avec cette technologie, ils ont pu charger un mur en une seule requête.

Enfin la partie cliente tire elle-aussi parti de la réactivité du backend servant les données du DDF en adoptant un paradigme SPA (Single Page Application). L'application web DDF consiste donc en un serveur web codé en NodeJS[15] et utilisant des composants d'interfaces basés sur la librairie ReactJS[16] et s'interfaçant avec l'API GraphQL grâce au client Apollo[17]. L'assemblage de ces technologies permet ainsi un chargement progressif des différents types de données nécessaires pour afficher par exemple tous les détails liés à une définition.

Conclusion

Le DDF a l'ambition de devenir un projet contributif incontournable dans le paysage de la francophonie en proposant un ensemble de ressources lexicographiques reflétant la grande diversité du français tel qu'il est parlé à travers le monde. A la fois base de données liées ouvertes, et plateforme contributive, le DDF se présente également comme un commun logiciel, dont le code source et les données sont disponibles sous licences libres[18] et exploitables à des fins de recherche. Son architecture repose sur le middleware générique et open-source Synaptix, développé par Mnémotix[19].

Le DDF est appelé à évoluer très prochainement en offrant de nouvelles fonctionnalités contributives qui permettront d'ajouter des localisations, des exemples, des marqueurs d'usages, et des relations sémantiques entre formes et ou définitions. Il sera également possible d'engager des discussions sur l'usage et l'étymologie des mots. Des espaces de discussion dédiés à la communauté des contributeurs seront également bientôt proposés pour soutenir son développement. De par sa nature résolument ouverte et inclusive, le DDF pose également des défis techniques et scientifiques quant à la curation et la modération des données contributives. Enfin une version du DDF capable de fonctionner hors-connexion est à l'étude et permettra d'élargir encore l'accès à ce nouveau bien commun des francophones.

[15] https://nodejs.org/
[16] https://fr.reactjs.org/
[17] https://www.apollographql.com/docs/react/
[18] voir https://gitlab.com/mnemotix/ddf pour le code source et https://www.dictionnairedesfrancophones.org/sparql pour le endpoint SPARQL
[19] middleware sémantique par ailleur au coeur d'autres communs logiciels reposant sur les technologies du web sémantique, comme notamment les outils développés par la coopérative https://www.elzeard.co/

Références

Dolar, K. (2017). Les dictionnaires collaboratifs en tant qu'objets discursifs, linguistiques et sociaux. PhD thesis. Université Paris Nanterre, Paris, France

Dolar, K., Steffens, M. & Gasparini, N. (2020). Dictionnaire des Francophones: A New Paradigm in Francophone Lexicography. In: Proceedings of XIX EURALEX Congress: Lexicography for Inclusion, Vol. I. Thrace: Democritus University of Thrace, pp. 23-30. https://euralex2020.gr/wp-content/uploads/2020/11/EURALEX2020_ProceedingsBook-p023-030.pdf

McCrae, J., Bosque-Gil, J., Garcia, J., Buitelaar, P. & Cimiano, P. (2017). The OntoLex-Lemon Model: development and applications. In Proceedings of eLex 2017, pp. 587-597. Accessed at https://elex.link/elex2017/wp-content/uploads/2017/09/paper36.pdf [30/05/2020]

Navarro, E., Sajous, F., Gaume, B., Prévot, L., Hsieh, S., Kuo, I., Magistry, P., et Huang, C.-R. (2009). Wiktionary and NLP: Improving synonymy networks. In Proceedings of the 2009 ACL-IJCNLP Workshop on The People's Web Meets NLP: Collaboratively Constructed Semantic Resources, 19–27, Suntec, Singapore. Association for Computational Linguistics

Steffens, M.., Dolar, K., & Gasparini, N. (2020). Structuration de données pour un dictionnaire collaboratif hybride. In: Terminologie & Ontologie: Théories et Applications. Actes de la conférence TOTh 2019. Chambéry: Presses Universitaires Savoie Mont Blanc, pp. 413-426.

Sajous, F., Hathout, N., et Calderone, B. (2013). GLÀFF, un Gros Lexique À tout Faire du Français. In Actes de la 20e conférence sur le Traitement Automatique des Langues Naturelles (TALN'2013), 285–298, Les Sables d'Olonne, France.

Sajous, F., Navarro, E., et Gaume, B. (2011). Enrichissement de lexiques sémantiques approvisionnés par les foules :le système WISIGOTH appliqué à Wiktionary. TAL, 52(1):11–35.

Sajous, F., Navarro, E., Gaume, B., Prévot, L., et Chudy, Y. (2010). Semi-automatic Endogenous Enrichment of Collaboratively Constructed Lexical Resources: Piggybacking onto Wiktionary. In Loftsson, H., Rögnvaldsson, E., et Helgadóttir, S. (eds), Advances in Natural Language Processing, vol. 6233 of LNCS, 332–344. Springer Berlin /eHeidelberg

Sérasset G. (2015). DBnary: Wiktionary as a Lemon-Based Multilingual Lexical Resource in RDF. Semantic Web – Interoperability, Usability, Applicability, IOS Press, 2015, Multilingual Linked Open Data, 6 (4), pp.355-361.